

COMPUTER METHOD AND APPARATUS FOR  
EXTRACTING DATA FROM WEB PAGES

ABSTRACT OF THE DISCLOSURE

Computer method and apparatus for extracting information from a Web page is  
5 disclosed. The invention apparatus is formed of an extractor coupled to receive Web  
pages from a source. The extractor uses natural language processing to extract desired  
information from the Web page. A storage subsystem receives from the extractor the  
extracted desired information and stores the extracted desired information in a database.  
The invention method for extracting data from a Web page includes the computer  
10 implemented steps of (i) using natural language processing, finding possible formal  
names on a given Web page, (ii) using pattern matching, searching the given Web page  
for formal names not found by the natural language processing, and (iii) refining a  
combined set of the found formal names to produce a working set of people and  
organization names extracted from the given Web page. The refining includes  
15 determining aliases of respective people and organization names, so as to effectively  
reduce duplicate names.